SNPFileCreator Documentation

Description:	Performs normalization and probe-level summarization to generate a SNP file for a set of Affymetrix SNP chip CEL files.
Author:	Marc-Danie Nazaire (Broad Institute), Joshua Gould (Broad Institute), David Twomey, gp-help@broadinstitute.org
Version:	4

Summary

The SNPFileCreator module accepts a ZIP archive containing CEL files generated using an Affymetrix SNP chip and, optionally, TXT files with SNP genotype call data. It can be set to perform probe intensity normalization using Quantile Normalization, and probe-level summarization using the Median Polish (default), Average Difference or Perfect Match/Mismatch (PM/MM) Difference Model (dChip) algorithms. Normalization, summarization, and association with the genotype call data accomplishes the preprocessing usually used for SNP array data.

The module generates a GenePattern BSNP or SNP file, which contains a matrix of intensity values per probe set and can be used as input for the other SNP analysis modules in GenePattern. (Note that this version of SNPFileCreator currently produces only allele-specific .bsnp and .snp files which are not accepted by some GenePattern SNP modules, such as LOHPaired version 3 and XChromosomeCorrect version 3)

Supported SNP Chips

SNPFileCreator supports CEL files generated using the following chips:

- Affymetrix Genome-Wide Human SNP Array 5.0 and 6.0
- Human Mapping 100K Set
- Human Mapping 500K Array Set
- Mapping 10K 2.0 Array Set
- 250K Array

The module accepts CEL file formats generated by either the Affymetrix® GeneChip® Command Console® Software (AGCC) or Affymetrix GeneChip® Operating Software (GCOS).

Memory Requirement

SNPFileCreator requires on average a minimum of 6GB RAM to process SNP Array 6.0 files. Use the GenePattern public server or configure your local GenePattern server to allocate the required memory for the module. Note that this module is currently only supported on 64-bit Linux servers and is therefore not available for download in the GenePattern Module Repository. To install this module on your local server, please contact the GenePattern team. (gp-help (at) broadinstitute.org)

References

1. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*. 2001;2:research0032. (http://genomebiology.com/2001/2/8/research/0032)
2. Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci USA*. 2007;104:20007-20012. (PDF: http://www.pnas.org/cgi/reprint/0710052104v1)

Parameters

Name	Description
cel.files	A ZIP archive that contains the CEL files and, optionally, the associated .txt genotype call files. The ZIP archive must be transferred to the GenePattern server before the SNPFileCreator module can begin processing. Due to the large size of the CEL files, this file transfer may take a considerable amount of time. Please be patient. Provide a file location or a URL to upload the ZIP archive; you can also provide a directory if you are running a local GenePattern server. **Note:** The CEL files must match the Chip Description File (CDF) used; this is particularly important if you upload a custom CDF. The genotype call files are in TXT format, associated with a specific sample, and contain 2 columns of data: the SNP IDs and the genotype calls.
genome annotation	A genome annotation file provided with SNPFileCreator. Options are UCSC hg19 and UCSC hg18. The counting of the chromosome locations in the provided genome annotations starts with 1 and not 0. You **must** select one of these files OR upload a custom *genome annotation file*.
genome annotation file	Custom genome information file. The annotation file must be tab delimited and contain the following three columns: Probe Set ID, Chromosome, and Physical Position. The counting of the chromosome locations in the genome annotations should start with 1 and not 0. You **must** select one of the genome annotation files provided with SNPFileCreator OR upload a custom *genome annotation file* here.

normalization.method	Whether to use a normalization method to compensate for systematic differences across chips (required): • Quantile Normalization (dChip) (default). This method takes the pooled distribution of probes on all chips. Then, for each value, it computes the quantile of that value in the distribution of probe intensities and transforms the original value to that quantile's value on the reference chip. This can also reduce noise among replicate measures of the same samples. • None **Note:** If you need the Invariant Set Normalization method to reproduce an older analysis, you will need to use an earlier version of SNPFileCreator, available on the Broad public server or in the module repository.
summarization.method	Method used to determine the summarized intensity value for each SNP based on the intensity levels of the probes in each probe set (required): • Median Polish (default): an iterative method that determines the significance of the values by examining the data gridded into a two-way table; it is resistant to anomalously high or low values • Average Difference: calculates the average difference between the PM probes and the MM probes in a probe set; if there are only PM probes on the array, selecting Average Difference will calculate just the average of the PM probes. • PM/MM Difference Model (dChip): the method dChip uses, and is described by Li and Wong (2001) NOTE: Median polish summarization is always applied if possible, otherwise average difference summarization, to all probe sets accept for SNP_XXX and CN_XXX probe sets.
sort.by.chromosome and location	Whether to sort the SNPs in the output file (required): • Do no sort • Sort (default); this sorts SNPs by chromosome and physical location. Probe sets that do not have location information are omitted from the output file. **Tip:** If you want to view your SNPs in the Integrative Genomics Viewer (IGV), select *Sort*, as IGV requires that the SNPs be sorted.

include.RandomGC.probesets	Whether to include RandomGC probe sets in the output file. The default (no) is to remove the RandomGC probe sets; this is the recommended setting. (required)
cdf.file	Use a custom Chip Description File (CDF), which describes the design of a chip, providing sequence and type information for the probes representing each of the probe sets on the chip. To use default CDF files provided by SNPFileCreator, leave this value blank. To view the list of CDF files provided by SNPFileCreator, in GenePattern, click the module and then click the *properties* link just below the version number. The files are listed in the *Current files* field of the properties form. (optional)
use.mismatch.probes	Whether to use mismatch (MM) probes in the summarization step (required): • yes • no (default) **Note:** If your data are older than 250K, you will probably want to include the mismatch probes.
output format	Determines the format in which the data will be output: • bsnp (default) • snp Note that bsnp is the required format for the Birdseed modules.
output.file	Base name for the output files (required). This file will have the extension .bsnp or .snp depending on the output file format

Input Files

1. ZIP archive containing the CEL files generated as Affymetrix output (required) and TXT files containing the genotype calls (optional). The genotype call file has two columns: the SNP ID and the call.
2. Chip Description File (CDF) (optional). The CDF file describes the design of a chip, providing sequence and type information for the probes representing each of the probe sets on the chip. Examples of Affymetrix CDF files can be found on the Affymetrix website.
3. Genome annotation file (required). This file contains chromosome number, transcription starting and ending sites (direction is from p-arm to q-arm for the human genome), and strand indication (a gene's sense strand relative to the genome sequence). Either one of the provided genome annotation files must be selected or a custom genome annotation file must be uploaded. Alternative annotation files can

be found on the Affymetrix website and other sources, but annotation files can also be made by hand. The annotation only needs to contain three columns: Probe Set ID, Chromosome, and Physical Position.

Output Files

1. stats.txt file (median, maximum, minimum, and 25th, 50th, and 75th percentile SNP intensity values for each sample)
2. BSNP or SNP file (Allele-Specific: contains raw intensity values for each allele per probe set. BSNP format is two rows per sample (alleles A & B), intensity values in columns. The SNP format is three columns per sample: intensity value for allele A, intensity value for allele B, and call); the SNPs in this file will be relative to an initial base of 1

Example Data

The example data is a subset of the data described in Beroukhim et al. (2007):

Input Parameter	Value
cel files genotype call files	ftp://ftp.broadinstitute.org/pub/genepattern/datasets/gistic/GISTIC_Hind_subset.zip

Platform Dependencies

Module type: SNP Analysis

CPU type: any

OS: any

Language: Java (minimum 1.5)

GenePattern Module Version Notes

Version	Description
V 4	SNPFileCreator module version 4 now supports Affymetrix Genome-Wide Human SNP Array 6.0 and contains improvements and bug fixes as follows: • Trimmed Mean and Median Probe have been removed from this version. When selected as summarization methods, they were causing the module to output values for the A alleles only.